



(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
28.11.2007 Bulletin 2007/48

(51) Int Cl.:
H01L 21/28 ^(2006.01) **H01L 21/8246** ^(2006.01)
H01L 27/115 ^(2006.01) **H01L 29/792** ^(2006.01)
G11C 16/04 ^(2006.01)

(21) Application number: **07252039.8**

(22) Date of filing: **18.05.2007**

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IS IT LI LT LU LV MC MT NL PL PT RO SE SI SK TR
Designated Extension States:
AL BA HR MK YU

(72) Inventors:
• **Lue, Hang-Ting**
Hsinchu (TW)
• **Lien, Hao-Ming**
Hsinchu (TW)

(30) Priority: **23.05.2006 US 419977**

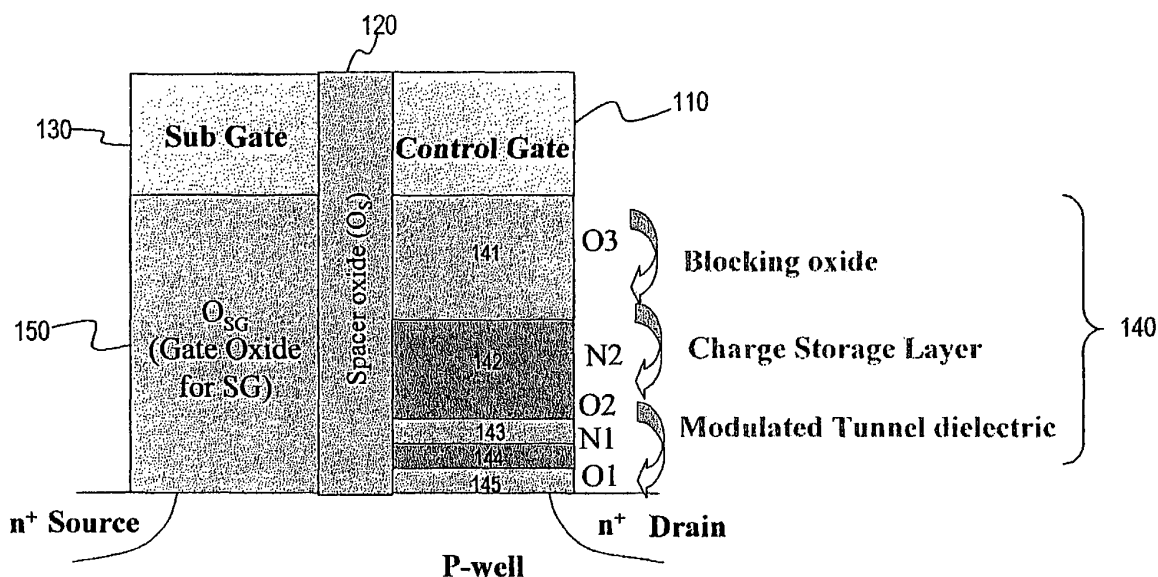
(74) Representative: **Johnson, Terence Leslie**
Marks & Clerk
90 Long Acre
London, WC2E 9RA (GB)

(71) Applicant: **MACRONIX INTERNATIONAL CO., LTD.**
Hsinchu (TW)

(54) **SONOS memory device**

(57) A bandgap engineered SONOS device structure for design with various AND architectures to perform a source side injection programming method. The BE-SONOS device structure (100) comprises a spacer oxide (O_s) disposed between a control gate (110) overlaying an oxide-nitride-oxide-nitride-oxide stack and a sub-gate (130) overlaying a gate oxide. In a first embodiment, a

BE-SONOS sub-gate-AND array architecture is constructed multiple columns of SONONOS devices with sub-gate lines and diffusion bitlines. In a second embodiment, a BE-SONOS sub-gate-inversion-bitline-AND architecture is constructed multiple columns of SONONOS devices with sub-gate inversion bitlines and with no diffusion bitlines.



100

Fig. 1

Description

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present invention relates generally to non-volatile memory devices, and more particularly, to nitride-based trapping-storage flash memories.

Description of Related Art

[0002] Electrically programmable and erasable non-volatile memory technologies based on charge storage structures known as Electrically Erasable Programmable Read-Only Memory (EEPROM) and flash memory are used in a variety of modern applications. A flash memory is designed with an array of memory cells that can be independently programmed and read. Sense amplifiers in a flash memory are used to determine the data value or values stored in a non-volatile memory. In a typical sensing scheme, an electrical current through the memory cell being sensed is compared to a reference current by a current sense amplifier.

[0003] A number of memory cell structures are used for EEPROM and flash memory. As the dimensions of integrated circuits shrink, greater interest is arising for memory cell structures based on charge trapping dielectric layers, because of the scalability and simplicity of the manufacturing processes. Memory cell structures based on charge trapping dielectric layers include structures known by the industry names Nitride Read-Only Memory, Poly-Silicon-Oxide-Nitride-Oxide-Silicon (SONOS), and PHINES, for example. These memory cell structures store data by trapping charge in a charge trapping dielectric layer, such as silicon nitride. As negative charge is trapped, the threshold voltage of the memory cell increases. The threshold voltage of the memory cell is reduced by removing negative charge from the charge trapping layer.

[0004] Nitride Read-Only Memory devices use a relatively thick bottom oxide, e.g. greater than 3 nanometers, and typically about 5 to 9 nanometers, to prevent charge loss. Instead of direct tunneling, band-to-band tunneling induced hot hole injection BTBTHH can be used to erase the cell. However, the hot hole injection causes oxide damage, leading to charge loss in the high threshold cell and charge gain in the low threshold cell. Moreover, the erase time must be increased gradually during program and erase cycling due to the hard-to-erase accumulation of charge in the charge trapping structure. This accumulation of charge occurs because the hole injection point and electron injection point do not coincide with each other, and some electrons remain after the erase pulse. In addition, during the sector erase of an Nitride Read-Only Memory flash memory device, the erase speed for each cell is different because of process variations (such as channel length variation). This difference in erase

speed results in a large V_t distribution of the erase state, where some of the cells become hard to erase and some of them are over-erased. Thus the target threshold V_t window is closed after many program and erase cycles and poor endurance is observed. This phenomenon will become more serious when the technology keeps scaling down.

[0005] A typical flash memory cell structure positions a tunnel oxide layer between a conducting polysilicon tunnel oxide layer and a crystalline silicon semiconductor substrate. The substrate refers to a source region and a drain region separated by an underlying channel region. A flash memory read can be executed by a drain sensing or a source sensing. For source side sensing, one or more source lines are coupled to source regions of memory cells for reading current from a particular memory cell in a memory array.

[0006] A traditional floating gate device stores 1 bit of charge in a conductive floating gate. The advent of Nitride Read-Only Memory cells in which each cell provides 2 bits of flash cells that store charge in an Oxide-Nitride-Oxide (ONO) dielectric. In a typical structure of a Nitride Read-Only Memory memory cell, a nitride layer is used as a trapping material positioned between a top oxide layer and a bottom oxide layer. The ONO layer structure effectively replaces the gate dielectric in floating gate devices. The charge in the ONO dielectric with a nitride layer may be either trapped on the left side or the right side of a NROM cell.

[0007] Floating gate devices encounter substantial scaling challenges due to inter-floating gate coupling, while nitride trapping device is flexible from such limitations. There are two main types of nitride trapping device: NROM that stores charges locally and SONOS that uses channel program/erase. These two types of devices have drawbacks. A Nitride Read-Only Memory device is sensitive to hot-hole induced damages, and a SONOS device suffers from retention problems caused by direct tunneling leakage through the thin tunnel oxide.

[0008] A conventional AND-type floating gate flash memory is suitable for many commercial applications because the memory device possesses the characteristics of high-density, low-power and fast speed programming. However, due to the inter-floating gate coupling effect, the scaling of AND-type floating gate devices is limited. When the space parameter for the floating gate device is shrunk, a high floating gate coupling effect may cause undesirable and severe disturbance. The conventional AND-type floating gate device also suffers from tunnel oxide scaling issues and erratic bits where a local defect, or trapped charge, in a tunnel oxide can result in the leakage of the charge in the floating gate.

[0009] To address the scaling issue in floating gate devices, charge trapping devices such as SONOS, MNOS or nano-crystal trapping devices are suggested. However, these devices all suffer serious charge retention problems. For a SONOS device, the ultra-thin tunnel oxide is unable to properly preserve a charge storage.

For a MNOS device, the structure does not provide a top oxide to block the charge loss. A nano-crystal device cannot be well-controlled because of the randomly distributed nano particles.

[0010] Accordingly, it is desirable to design AND-type floating gate flash memories that provide scalability while overcoming the retention problems as well as maintaining efficient hole tunneling erase.

SUMMARY OF THE INVENTION

[0011] The present invention provides a bandgap engineered SONOS (referred to as "BE-SONOS" or "SONONOS") device structure for design with various AND architectures to perform a source side injection (SSI) programming method. The BE-SONOS device structure comprises a spacer oxide disposed between a control gate overlaying an oxide-nitride-oxide-nitride-oxide ($O_3-N_2-O_2-N_1-O_1$) stack and a sub-gate overlaying a gate oxide. In a first embodiment, a BE-SONOS SG-AND (sub-gate-AND) array architecture is constructed multiple columns of SONONOS devices with sub-gate lines and diffusion bitlines. In a second embodiment, a BE-SONOS SGIB-AND (sub-gate-inversion-bitline-AND) architecture is constructed multiple columns of SONONOS devices with sub-gate inversion bitlines and with no diffusion bitlines.

[0012] Broadly state, an integrated circuit device comprises a semiconductor substrate; a plurality of memory cells on the semiconductor sub-trate, each memory cell having a spacer oxide disposed between a gate and a sub-gate, each gate overlaying a blocking oxide-charge storage layer-modulated tunnel dielectric stack, each sub-gate overlaying a gate oxide; and an N+ buried diffusion disposed in the semiconductor substrate and positioned underneath between a first gate oxide and a first blocking oxide-charge storage layer-modulated tunnel dielectric stack that serves as a first diffusion bitline.

[0013] Advantageously, the BE-SONOS AND array architectures of the present invention provides greater scalability over floating gate and AND type of memory devices. The present invention also advantageously provides a uniform and self-converging channel hole tunneling erase operation. Moreover, the present invention eliminates the inter-floating gate coupling effect. The present invention further provides desirable reliability properties including predictable characteristic of excellent charge retention, predictable number of nearly no erratic bits, and predicable small degradation after program and erase cycles.

[0014] The structures and methods regarding to the present invention are disclosed in the detailed description below. This summary does not purport to define the invention. The invention is defined by the claims. These and other embodiments, features, aspects, and advantages of the invention will become better understood with regard to the following description, appended claims and accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] Fig. 1 is a structural diagram illustrating a cross-sectional view of a single cell structure of an n-channel BE-SONOS device with sub-gate in accordance with the present invention.

[0016] Fig. 2A is a circuit diagram illustrating a first embodiment of a BE-SONOS SG-AND array architecture with diffusion bitlines in accordance with the present invention; Fig. 2B is a layout diagram illustrating the first embodiment of the BE-SONOS SG-AND array architecture with diffusion bitlines in accordance with the present invention.

[0017] Fig. 3A is a layout diagram illustrating a cross-sectional view in channel length direction of the BE-SONOS SG-AND array architecture in the first embodiment in accordance with the present invention; Fig. 3B is a layout diagram 350 illustrating a cross-sectional view in channel width direction of the BE-SONOS SG-AND array architecture in the first embodiment in accordance with the present invention.

[0018] Fig. 4A is a circuit diagram illustrating an electrical reset for the SONONOS SG-AND array architecture in the first embodiment in accordance with the present invention; Fig. 4B is a graphical diagram illustrating a waveform of a self-converging reset in the first embodiment in accordance with the present invention.

[0019] Fig. 5A is a circuit diagram illustrating an electrical program for the SONONOS SG-AND array architecture in the first embodiment in accordance with the present invention; Fig. 5B is a layout diagram illustrating the electrical program for the SONONOS SG-AND array architecture in the first embodiment in accordance with the present invention.

[0020] Fig. 6A is a circuit diagram illustrating an electrical erase for the SONONOS SG-AND array architecture of the first embodiment in accordance with the present invention; Fig. 6B is a graphical diagram illustrating a waveform of a self-converging erase in accordance with the present invention.

[0021] Fig. 7 is a circuit diagram illustrating a read operation for the SONONOS SG-AND array architecture in the first embodiment in accordance with the present invention.

[0022] Fig. 8A is a circuit diagram illustrating a second embodiment of a BE-SONOS SGIB-AND array architecture in accordance with the present invention; Fig. 8B is a layout diagram illustrating the second embodiment of the BE-SONOS SGIB-AND array architecture in accordance with the present invention.

[0023] Fig. 9A is a layout diagram illustrating the cross sectional view in channel length direction of the SONONOS SGIB-AND array architecture in the second embodiment in accordance with the present invention; Fig. 9B is a layout diagram illustrating the cross sectional view in channel width direction of the SONONOS SGIB-AND array architecture in the second embodiment in accordance with the present invention.

[0024] Fig. 10A is a circuit diagram illustrating an electrical reset of the SONONOS SGIB-AND array architecture in the second embodiment in accordance with the present invention; Fig. 10B is a graphical diagram illustrating a waveform of a self-converging reset in accordance with the present invention.

[0025] Fig. 11A is a circuit diagram illustrating an electrical program of the SONONOS SGIB-AND array architecture in the second embodiment in accordance with the present invention; Fig. 11B is a layout diagram illustrating the electrical program for the SONONOS SGIB-AND array architecture in the second embodiment in accordance with the present invention.

[0026] Fig. 12A is a circuit diagram illustrating an electrical erase of the BE-SONOS SGIB-AND array architecture of the second embodiment in accordance with the present invention; Fig. 12B is a graph diagram illustrating a waveform of a self-converging erase of the second embodiment in accordance with the present invention.

[0027] Fig. 13A is a circuit diagram illustrating a read operation of the SONONOS SGIB-AND array architecture in the second embodiment in accordance with the present invention; Fig. 13B is a layout diagram illustrating the read operation of the SONONOS SGIB-AND array architecture of the second embodiment in accordance with the present invention.

DETAILED DESCRIPTION OF EMBODIMENTS OF THE INVENTION

[0028] Referring now to Fig. 1, there is shown a structural diagram illustrating a cross-sectional view of a single cell structure of an n-channel BE-SONOS device 100 with sub-gate (SG). The n-channel BE-SONOS device 100 includes a spacer oxide (O_s) 120 that is disposed between a control gate 110 and a sub-gate 130. An oxide-nitride-oxide-nitride-oxide ($O_3-N_2-O_2-N_1-O_1$) structure 140 is disposed underneath the control gate 110. A gate oxide O_{SG} 150 is disposed underneath the sub-gate 130. The $O_3-N_2-O_2-N_1-O_1$ structure 140 includes a blocking oxide O_3 141, a charge storage layer N_2 142, and a modulated tunnel dielectric $O_2-N_1-O_1$ 143-145. The bottom $O_2-N_1-O_1$ 143-145 layers provide hole tunneling current and good data retention.

[0029] The n-channel BE-SONOS device 100 is a five-terminal device with two gates, the control gate 110 and the sub-gate 130. Underneath the control gate 110, there is the $O_3-N_2-O_2-N_1-O_1$ structure 140 for the charge storage. Underneath the sub-gate 130, there is the non-trapping gate oxide 150. The control gate 110 can control program, erase, and read the charge storage layer. The sub-gate 130 can provide source side injection (SSI) programming method. The source side injection is a low-power and high-speed programming method. The $O_1-N_1-O_2$ layer can be implemented with ultra-thin oxide and nitride, typically within 3 nm to provide hole direct tunneling. The N_2 layer 142 is thicker than 5 nm to provide

higher trapping efficiency. In the layer 141 formation method, one technique is to use a wet converted top oxide to provide a large density of traps at the interface between O_3 and N_2 . The O_3 layer is typically thicker than 6 nm in order to prevent charge loss from top oxide. The $O_1-N_1-O_2$ layers serve as a tunneling dielectric for the hole tunneling.

[0030] An exemplary set of device parameters for the n-channel BE-SONOS device 100 with the sub-gate 130 is shown below.

Bottom Oxide (O_1)	15 Å
Inter Nitride (N_1)	20 Å
Inter Oxide (O_2)	18 Å
Trapping Nitride (N_2)	90 Å
Gate Oxide for SG (O_{SG})	150 Å
Spacer Oxide (O_s)	200 Å
Gate material	N+ - poly or P+ - poly gate

[0031] In Fig. 2A, there is shown a circuit diagram illustrating a first embodiment of a BE-SONOS SG-AND array architecture 200 with diffusion bitlines. A plurality of SONONOS devices are connected in parallel to form the BE-SONOS SG-AND array architecture 200. The BE-SONOS array architecture 200 comprises a plurality of wordlines WL0 210, WL1 211, WL2 212, WLm 213 intersecting a plurality of bitlines BL0 220, BL1 221, BL2 222, BL3 223, BL4 224 and BLn 225. A corresponding sub-gate line is parallel and located nearby a bitline. A sub-gate SG1 230 is located adjacent to the bitline BL0 220. A sub-gate SG2 231 is located adjacent to the bitline BL1 221. A sub-gate SG3 232 is located adjacent to the bitline BL2 222. A sub-gate SG4 233 is located adjacent to the bitline BL3 223. A sub-gate SG5 234 is located adjacent to the bitline BL4 224. A sub-gate SGn 235 is located adjacent to the bitline BL5 225. A sample BE-SONOS (or SONONOS) device 240 that functions as a memory cell is shown in a circled area.

[0032] As shown in Fig. 2B, there is a layout diagram 250 illustrating the first embodiment of the BE-SONOS SG-AND array architecture 200 with diffusion bitlines. Although each sub-gate SG is in parallel with a corresponding bitline, each SG has a slight offset from the corresponding bitline, e.g. the SG1 230 is positioned slightly to the right of BL0 220. Every bitline can either serve as a source or a drain. Each SG is positioned between two bitlines, e.g. the SG1 230 between BL0 220 and BL1 221. Each bitline in the BL0 220, BL1 221, BL2 222, BL3 223, BL4 224, and BL5 225 can function as a source region or a drain region. Therefore, the SG1 230 is disposed between the source region in the BL0 220 and the drain region in the BL1 221. The parameters W 240 and Ws 242 are approximately equal to the parameter F, where the parameter F denotes the critical dimen-

sion in a technology node. For example, the parameter F is equal to 50 nm for a 50 nm node.

[0033] Fig. 3A is a layout diagram 300 illustrating a cross-sectional view in channel length direction of the BE-SONOS SG-AND array architecture in the first embodiment. The spacer oxide 120 separates the control gate 110 and the sub-gate 130. The $O_3-N_2-O_2-N_1-O_1$ structure 140 is disposed underneath the control gate 110. The gate oxide O_{SG} 150 is disposed underneath the gate 130. A suitable implementation of the control gate 110 is poly-1, and a suitable implementation of the sub-gate 130 is poly-2. N^+ buried diffusion (BD) wells 330, 332, 334 and 336 are implemented for diffusion bitlines (BLs). In the layout diagram 300, a first cell structure comprises the control gate 110 and the sub-gate 130, with an adjacent and second cell structure that comprises a gate 310 and a sub-gate 312, with an adjacent and third cell structure that comprises a gate 320 and a sub-gate 322.

[0034] In Fig. 3B, there is shown a layout diagram 350 illustrating a cross-sectional view in channel width direction of the BE-SONOS SG-AND array architecture in the first embodiment. The gap between the gate 310 and the gate 320 is denoted by a parameter Ws 360, which provides isolation between the gate 310 and the gate 320. Other similar isolations are shown between two gates to provide an isolation between two gates. The pitch in the channel width direction is approximately equal to $2F$ smaller than that in channel length direction $3F$ caused by a diffusion bitline. Therefore, the BE-SONOS SG-AND architecture is approximately equal to $6F^2$ per cell.

[0035] Fig. 4A is a circuit diagram 400 illustrating an electrical reset for the SONONOS SG-AND array architecture of the first embodiment. During the electrical reset, the wordlines (or gates) WL_0 210, WL_1 211, WL_2 212, and WL_m 213 are set to -10 volts, the bitlines BL_0 220, BL_1 221, BL_2 222, BL_3 223, BL_4 224 and BL_5 225 are left floating, and the sub-gates SG_1 230, SG_2 231, SG_3 232, SG_4 233, SG_5 234, and SG_n 235 are set to 0 volt. In one embodiment, the odd number sub-gates are electrically connected together, including SG_1 230, SG_3 232 and SG_5 234, while the even number sub-gates are electrically connected together, including SG_2 231, SG_4 233 and SG_n 235. Before operations, the memory circuit 400 is reset by applying $V_{gb} = -15V$ (or partition the gate voltage into each WL and p-well), which produces a desirable self-converging property, as shown in a graph 450 in Fig. 4B. Even if the BE-SONOS device is initially charged to various V_t , the reset operation can tighten these initial points to the reset/erase state. A typical reset time is around 100 msec. In one example, the n-channel BE-SONOS with $ONONO = 15/20/18/70/90$ Angstrom, and a N^+ -poly gate. $L_g/W = 0.22/0.16$ μm .

[0036] To state in another way, a reset operation is carried out to tighten the V_t distribution before operations. In contrast to a floating gate device where there is no self converging erase, the BE-SONOS provides a self-converging erase reset/erase methods, which is necessary

because the initial V_t distribution is often widely distributed due to the process issues, such as plasma charging effect. The self-converging reset assists in tighten the initial V_t distribution.

[0037] Fig. 5A is a circuit diagram 500 illustrating an electrical program for the SONONOS SG-AND array architecture in the first embodiment, while Fig. 5B is a layout diagram 550 illustrating the electrical program for the SONONOS SG-AND array architecture in the first embodiment. In one example during electrical program, the wordline WL_1 211 is set to 10 volts while the other wordlines WL_0 210, WL_2 212, and WL_m 213 are set 0 volt. The bitline BL_1 221 is set to 5 volts, and the bitlines BL_0 220, BL_2 222, BL_3 223, BL_4 224, and BL_n 225 are set to 0 volt. The odd number sub-gates SG_1 230, SG_3 232 and SG_5 234 are set to 1 volt, while the even number sub-gates SG_2 231, SG_4 233, and SG_n 235 are set to 0 volt. The bitlines BL_0 220, BL_1 221, BL_2 222, BL_3 223, BL_4 224 and BL_5 225 provide a greater flexibility in programming than sub-gates SG_1 230, SG_2 231, SG_3 232, SG_4 233, SG_5 234, and SG_n 235 because each bitline can be independently programmed, while the sub-gates are programmed based on the even number or odd number of sub-gates. One type of electrical programming method is a source side injection. The source side injection programs a cell to a high voltage threshold V_t state. For example, the source injection applies $V_g = 10V$ to the selected WL_1 , $V_g = 0V$ to other wordlines, $SG = 1V$ for programming, and $SG = 0V$ for inhibition. The voltage setting of the SG voltage at 1 volt is intended as an illustration such that in general it is typically 0.5 to 2 volts higher than the threshold voltage under SG gate.

[0038] When a cell-A 422 is selected programming, the SG is set to 1 volt so that the channel underneath SG is slightly turned on. Electrons are injected into the cell-A 422 by source side injection method to make the voltage threshold, V_t , higher than PV . The SG for a cell-B 424 is set to 0 volt, which turns SG off so that there is no injection into the cell-B 424. As for a cell-C 426, the SG is set to 1 volt where the $WL = 0$ volt which turns off cell-C 426 so that there is also no injection into the cell-C 426. As a result, programming can be randomly selected with adequate program inhibit technique.

[0039] To carry out an electrical program, the selected wordline is applied a high voltage, 10 volts, and the sub-gate is applied 1 volt to perform a source side injection. The source side injection is a low-power and high-speed programming method. One of skill in the art should recognize that parallel programming methods such as page programming with 2kB cells in parallel can burst the programming throughput to more than 10 MB per second while the cell current consumption can be controlled within 2 mA. To avoid program disturbance to other bitlines, the sub-gate SG_2 231 is set to 0 volt and turns off the inhibit cell.

[0040] As illustrated in Fig. 6A, there is a circuit diagram 600 showing an electrical erase for the SONONOS SG-AND array architecture in the first embodiment. The

erase operation is executed similar to the reset operation. During electrical erase, the wordlines WL0 210, WL1 211, WL2 212, and WLn 213 are set to -10 volts, the bitlines BL0 220, BL1 221, BL2 222, BL3 223, BL4 224 and BL5 225 are left floating, and the sub-gates SG1 230, SG2 231, SG3 232, SG4 233, SG5 234, and SGn 235 are set to 0 volt. The erase operation is performed in the unit of a sector or block. The BE-SONOS device produces desirable self-converging erase property, as shown in a graph 650 in Fig. 6B. The erase saturation V_t is dependent on the parameter V_g . A higher V_g causes a higher saturated V_t . The convergent time is typically around 10 to 100 msec.

[0041] Fig. 7 is a circuit diagram 700 illustrating a read operation for the SONONOS SG-AND array architecture in the first embodiment. In one example during a read operation of the cell-A 422, the wordline WL1 211 is set to 5 volts while the other wordlines WL0 210, WL2 212, and WLn 213 are set 0 volt. The bitline BL1 221 is set to 1 volt, and the bitlines BL0 220, BL2 222, BL3 223, BL4 224, and BLn 225 are set to 0 volt. The odd number sub-gates SG1 230, SG3 232 and SG5 234 are set to 3 volts, while the even number sub-gates SG2 231 and SG4 233 are set to 0 volt. A read operation is performed by applying a gate voltage that is between an erased state V_t (EV) and a programmed state V_t (PV). The gate voltage is typically around 5 volts. Alternatively, the gate voltage can be selected to be more than 5 volts or less than 5 volts, provided the gate voltage falls in the range of a high V_t value and a low V_t value. If V_t of the cell-A 422 is higher than 5 volts, then the read current is likely to be a very small value (e.g., $<0.1 \mu A$). If V_t of the cell-A 422 is less than 5 volts, the read current is likely to be a high value (e.g., $>0.1 \mu A$).

[0042] The applied voltage at a bitline (BL) is typically around 1 volt. A larger read voltage will induce more current, but the read disturbance may be larger. The WL number of SG-AND string is typically 64, 128, or 256. A larger number of SG-AND string may save more overhead and increase the array efficiency. However, the program distribution may be larger. A trade-off is weighed in choosing an adequate number of SG-AND string.

[0043] Although the above read function describes a random access read operation, one of ordinary skill in the art should recognize that a page read of multiple cells are possible without departing from the spirits of the present invention.

[0044] Turning now to Fig. 8A, there is shown a circuit diagram 800 illustrating a second embodiment a BE-SONOS (or SONONOS) SGIB-AND (Sub-gate Inversion Bitline-AND) array architecture, while Fig. 8B is a layout diagram 850 illustrating the second embodiment of the BE-SONOS SGIB-AND array architecture. The term SGIB means that a bitline is formed from an inversion layer by turning on a sub-gate. Unlike the first embodiment where the BE-SONOS structure as shown in the layout diagram 300 has a N+ buried diffusion, the SONOS SGIB-AND cell structure has no N+ buried dif-

fusion, as shown in Fig. 9, and therefore, there is no offset between a bitline and a sub-gate. The SONONOS devices are connected in parallel to form an AND array with SG in which there are no diffusion bitlines.

[0045] The BE-SONOS array architecture 800 comprises a plurality of wordlines WL0 810, WL1 811, WL2 812, WLn 813 intersecting a plurality of bitlines BL0 820, BL1 821, BL2 822, BL3 823 and BL4 824. A corresponding sub-gate line is parallel to each bitline. A sub-gate SG0 830 is placed in parallel to the bitline BL0 820. A sub-gate SG1 831 is placed in parallel to the bitline BL1 821. A sub-gate SG2 832 is placed in parallel to the bitline BL2 822. A sub-gate SG3 833 is placed in parallel to the bitline BL3 823. A sample BE-SONOS (or SONONOS) device 840 that functions as a memory cell is shown in a circled area.

[0046] In the SGIB-AND array architecture 850, every fourth sub-gates are commonly electrically connected, i.e. the SG0 830, SG4 834, SG8, etc. are electrically connected together, the SG1 831, SG5, SG9, etc are electrically connected together, the SG2 832, SG6, SG10, etc are electrically connected together, and the SG3 833, SG7, SG11, etc are electrically connected together.

[0047] As shown in the layout diagram 850 in Fig. 8B, there are no N+ region underneath each of the SG0 830, the SG1 831, the SG2 832, the SG3 833, the SG4 834, and the SG 835 in the memory array of the SGIB-AND architecture 800. The overall dimension of a cell size in the SGIB-AND architecture 800 is reduced relative to a cell size as shown in the SG-AND array architecture 200 in Fig. 2A.

[0048] When the SG0 830, the SG1 831, the SG2 832, the SG3 833, the SG4 834, and the SG 835 are turned on, each creates an N-channel inversion layer, which effectively serves as a barrier diffusion layer to function, respectively, as a source/drain 860, a source/drain 861, a source/drain 862, a source/drain 863, a source/drain 864 and a source/drain 865. Each sub-gate in the SG0 830, the SG1 831, the SG2 832, the SG3 833, the SG4 834, and the SG 835 therefore serves dual functions. The first function that each sub-gate in the SG0 830, the SG1 831, the SG2 832, the SG3 833, the SG4 834, and the SG 835 serves is a sub-gate for a source side injection programming. The second function that each sub-gate in the SG0 830, the SG1 831, the SG2 832, the SG3 833, the SG4 834, and the SG 835 serves is an inversion bitline when a subgate is turned on. Each source/drain in the source/drain 860, the source/drain 861, the source/drain 862, the source/drain 863, the source/drain 864, and the source/drain 865 is for connecting to a metal bitline. The symbol L_g 870 denotes the drawn channel length. The symbol W 874 denotes the channel width. Typically, the parameters W 874, W_s 876, L_g 870, L_s 872 are approximately equal to the parameter F , where the parameter F represents the critical dimension in a technology node. For example, the parameter F is equal to 50 nm for a 50 nm node.

[0049] As shown in Fig. 9A, there is a layout diagram

900 illustrating the cross sectional view in channel length direction of the SONOS SGIB-AND array architecture in the second embodiment, while Fig. 9B is a layout diagram illustrating the cross sectional view in channel width direction of the SONOS SGIB-AND array architecture in the second embodiment. Each subgate SG in SG 910, SG 912, SG 914, SG 916 and SG918 has no implant of source region or drain region, which means that there is no N⁺ region underneath each subgate SG in SG 910, SG 912, SG 914, SG 916 and SG 918. The pitch in a channel width direction is approximately equal to that in a channel length direction 2F. Therefore, the BE-SONOS SGIB-AND architecture is approximately equal to 4F² per cell.

[0050] In Fig. 10A, there is a circuit diagram illustrating an electrical reset of the SONOS SGIB-AND array architecture 1000 in the second embodiment. During the electrical reset, the wordlines (or gates) WL0 810, WL1 811, WL2 812, and W-Lm 813 are set to -10 volts, the bitlines BL0 820, BL 1 821, BL2 822, BL3 823, BL4 824 and BL5 825 are left floating, the sub-gates SG0 830, SG1 831, SG2 832, SG3 833, SG4 834, and SGn 835 are set to 0 volt, and P-well 1010 is set to 5 volts. When a sub-gate SG = is equal to zero volt, that means the SG is not turned on so that there is no inversion bitline. In one embodiment, every fourth sub-gates are connected together such that SG0 830 is connected to SG4, SG1 831 is connected to SG5, and so on.

[0051] Before operations, the memory circuit 1000 is reset by applying V_{gb} = -15V (or partition the gate voltage into each WL and p-well), which produces a desirable self-converging property, as shown in the graph 1050 in Fig. 10B. Various circles and triangles in the graph 1050 represent different initial points over a wide distribution where these points converges to a threshold voltage V_t. Even if the BE-SONOS device is initially charged to various V_t, the reset operation can tighten these initial points to the reset/erase state. A typical reset time is around 100 msec. In one example, the n-channel BE-SONOS device with ONONO = 15/20/18/70/90 Angstrom, and a N⁺-poly gate. L_g/W = 0.22/0.16 μ m.

[0052] To phrase in another way, a reset operation is carried out to tighten the V_t distribution before operations. In contrast to a floating gate device where there is no self-converging erase, the BE-SONOS provides a self-converging erase reset/erase methods, which is necessary because the initial V_t distribution is often widely distributed due to the process issues, such as plasma charging effect. The self-converging reset assists in tighten the initial V_t distribution.

[0053] As illustrated in Fig. 11A, there is a circuit diagram illustrating an electrical program of the SONOS SGIB-AND array architecture 1100 in the second embodiment, while Fig. 11B is a layout diagram 1150 illustrating the electrical program for the SONOS SGIB-AND array architecture in the second embodiment. In one example during electrical program of a cell A1110, the wordline WL1 811 is set to 10 volts while the other

wordlines WL0 810, WL2 812, and W-Lm 813 are set 0 volt. The bitlines BL0 820, BL1 821, and BL3 823 are left floating. The bitline BL2 is set to 0 volt, and the bitline BL4 is set to 5 volts. The sub-gates SG0 830 and SG4 834 are set to 8 volts. The sub-gate SG1 is set to 0 volt, the sub-gate SG2 is set to 5 volts, and the sub-gate SG3 is set to 1 volt.

[0054] The bitlines BL0 820, BL1 821, BL2 822, BL3 823, and BL4 824 provide a greater flexibility in programming than sub-gates SG0 830, SG1 831, SG2 832, SG3 833 and SG4 834 because each bitline can be independently programmed. One type of electrical programming method is a source side injection. The source side injection programs a cell to a high voltage threshold V_t state. For example, the source injection applies V_g = 10 V to the selected WL1, V_g = 0V to other WL's, SG = 1V for programming and SG = 0V for inhibition. The voltage setting of the SG voltage at 1 volt is intended as an illustration such that in general it is 0.5 to 2 volts higher than the threshold voltage under SG gate.

[0055] To carry out an electrical program, the selected wordline is applied a high voltage, 10 volts, and the sub-gate SG3 833 is applied 1 volt to perform a source side injection to program a target cell. The SG1 831 is set to 0 volt for program inhibit and the SG4 834 is set to 8 volts to provide sufficient overdrive to reduce a bitline resistance. One of skill in the art should recognize that parallel programming methods such as page programming with 2kB cells in parallel can burst the programming throughput to more than 10 MB per second while the cell current consumption can be controlled within 2 mA. To avoid program disturbance to other bitlines, the sub-gate SG2 231 is set to 0 volt and turns off the inhibit cell.

[0056] The electrical programming is to conduct a source side injection to program a cell to a high voltage threshold, V_t, state. For example in the electrical programming of the cell A 1110, the operations apply V_g = 10 V to the selected WL₁ 811, apply V_g = 0V to other wordlines including WL0 810, WL2 812 and W-Lm 813, set SG3= 1V for programming, set SG1 = 0V for program inhibition, and set SG2=5V for pass gate. The sub-gate SG4 834 is set to 8 volt to highly turn on the sub-gate SG4 834 so that the inversion layer potential can be raised up to 5 volts. In programming the cell A 1110, the sub-gate SG3 833 is set to 1 volt so that the source side injection occurs.

The threshold voltage V_t is raised to above the programming voltage, P_V. A program inhibition is provided to a cell-B 1012, a cell-C 1014 and a cell-D 1016.

[0057] Fig. 12A is a circuit diagram 1200 illustrating an electrical erase of the BE-SONOS SGIB-AND array architecture in the second embodiment, while Fig. 12B is a graph diagram 1250 illustrating a desirable self-converging erase property with respect to the second embodiment. The erase operation is similar to the reset operation. During electrical erase, the wordlines WL0 810, WL1 811, WL2 812, and W-Lm 813 are set to -10 volts, the bitlines BL0 820, BL1 821, BL2 822, BL3 823, BL4

824 and BL5 825 are left floating, and the sub-gates SG2 830, SG1 831, SG2 832, SG3 833, SG4 834 are set to 0 volt. The electrical erase is performed in the unit of a sector or block. The BE-SONOS device produces a desirable self-converging erase property, as shown in a graph diagram 1250 in Fig. 6B. The erase saturation V_t is dependent on V_g . A higher V_g causes a higher saturated V_t . The convergent time is typically around 10 to 100 msec.

[0058] Fig. 13A is a circuit diagram 1300 illustrating a read operation of the SONONOS SGIB-AND array architecture in the second embodiment, while Fig. 13B is a layout diagram 1350 illustrating the read operation of the SONONOS SGIB-AND array architecture in the second embodiment. In one example during a read operation of the cell A 1110, the wordline WL1 811 is set to 5 volts while the other wordlines WL0 810, WL2 812, and WLM 813 are set 0 volt. The bitlines BL0 820, BL1 821, and FL4 824 are left floating. The bitline BL2 822 is set to 0 volt and the bitline BL3 823 is set to 1 volt. The sub-gates SG0 830, SG1 831, and SG4 834 are set to 0 volts, while SG2 832 and SG3 833 are set to 5 volts. A read operation is performed by applying a gate voltage that is between an erased state V_t (EV) and a programmed state V_t (PV). The gate voltage is typically around 5 volts. Alternatively, the gate voltage can be selected to be more than 5 volts or less than 5 volts, provided the gate voltage falls in the range of a high V_t value and a low V_t value. If V_t of the cell-A 422 is higher than 5 volts, then the read current is likely to be a very small value (e.g., $<0.1 \mu A$). If V_t of the cell-A 422 is less than 5 volts, the read current is likely to be a high value (e.g., $>0.1 \mu A$). The state of a memory can then be identified.

[0059] The applied voltage at a bitline (BL) is typically around 1 volt. A larger read voltage will induce more current, but the read disturbance may be larger. The WL number of SG-AND string is typically 64, 128, or 256. A larger number of SG-AND string may save more overhead and increase the array efficiency. However, the program distribution may be larger. A trade-off is weighed in choosing an adequate number of SGIB-AND string.

[0060] Although the above read function describes a random access read operation, one of ordinary skill in the art should recognize that a page read of multiple cells are possible without departing from the spirits of the present invention. The invention has been described with reference to specific exemplary embodiments. Various modifications, adaptations, and changes may be made without departing from the spirit and scope of the invention. Accordingly, the specification and drawings are to be regarded as illustrative of the principles of this invention rather than restrictive, the invention is defined by the following appended claims.

Claims

1. An integrated circuit device comprising:

a semiconductor substrate;
a plurality of memory cells on the semiconductor substrate, each memory cell having a spacer dielectric layer disposed between a gate and a sub-gate, each gate overlaying a blocking oxide-charge storage layer-modulated tunnel dielectric stack, each sub-gate overlaying a gate oxide; and
an N+ buried diffusion disposed in the semiconductor substrate and positioned underneath between a first gate oxide and a first blocking oxide-charge storage layer-modulated tunnel dielectric stack that serves as a first diffusion bitline.

2. An integrated circuit according to claim 1, wherein the modulated tunnel dielectric stack comprises an oxide-nitride-oxide ($O_2-N_1-O_1$) stack.

3. An integrated circuit according to claim 2, wherein the ($O_2-N_1-O_1$) stack comprises a substantially thin oxide and nitride layer.

4. An integrated circuit device comprising:

a semiconductor substrate; and
a plurality of memory cells on the semiconductor substrate, each memory cell having a spacer oxide disposed between a gate and a sub-gate, each gate overlaying a blocking oxide-charge storage layer-modulated tunnel dielectric stack, each sub-gate overlaying a gate oxide;

wherein a first sub-gate serving as source side injection during programming when the first sub-gate is a slightly turn-on state; and
wherein the first sub-gate serving as an inversion layer for connection to a first metal bitline when the first sub-gate is an ON state.

5. An integrated circuit according to claim 4, wherein the modulated tunnel dielectric stack comprises an oxide-nitride-oxide ($O_2-N_1-O_1$) stack.

6. An integrated circuit according to claim 5, wherein the ($O_2-N_1-O_1$) stack comprises a substantially thin oxide and nitride layer.

7. An SG-AND array architecture comprising:

a memory array having a plurality of columns of SONONOS devices that are connected in parallel;
a plurality of sub-gate lines, each sub-gate line connecting to a corresponding column of SONONOS devices; and
a plurality of diffusion bitlines, each diffusion bitline connecting to a corresponding column of SONONOS devices.

8. An SG-AND array architecture according to claim 7, wherein the plurality of sub-gate lines are used for source side injection during programming.
9. An SG-AND array architecture according to claim 7, wherein the odd number of sub-gate lines are commonly electrically connected.
10. An SG-AND array architecture according to claim 7, wherein the even number of sub-gate lines are commonly electrically connected.
11. An SGIB-AND array structure comprising:
- a memory array having a plurality of columns of SONOS devices that are connected in parallel structure, the memory array having no diffusion bitlines; and
 - a plurality of sub-gate lines, each sub-gate line connecting to a corresponding column of SONOS devices.
12. An SGIB-AND array architecture according to claim 11, wherein the plurality of sub-gate lines are used for source side injection during programming when the plurality of sub-gate lines are in an off state.
13. An SGIB-AND array architecture according to claim 11, wherein each of the plurality of sub-gate lines functions as inversion layers for connection to a respective metal bitline when the plurality of sub-gate lines are in an ON state.
14. An SGIB-AND array architecture according to claim 11, wherein the plurality of sub-gate lines in which every Nth sub-gate lines are commonly electrically connected.
15. An SGIB-AND array architecture according to claim 11, wherein the plurality of sub-gate lines in which every 4th sub-gate lines are commonly electrically connected.
16. A memory cell structure comprising:
- a spacer oxide having a first sidewall and a second sidewall;
 - a sub-gate extending horizontally to the first sidewall of the spacer oxide;
 - an oxide-nitride-oxide ($O_2-N_1-O_1$) stack extending horizontally to the second sidewall of the spacer oxide; and
 - a control gate overlaying a third oxide layer O_3 and extending horizontally to the second side wall of the spacer oxide.
17. A memory structure according to claim 16, further comprising a gate oxide underlying the sub-gate and extending horizontally to the first sidewall of the spacer oxide.
18. A memory structure according to claim 17, further comprising a second nitride layer N_2 disposed between the control gate and the second oxide layer O_2 , the second nitride layer N_2 extending horizontally to the second sidewall of the spacer oxide, the second nitride layer N_2 serving as a charge storage layer.
19. A memory structure according to claim 18, further comprising a third oxide layer O_3 disposed between the control gate and the second nitride layer N_2 , the third oxide layer extending horizontally to the second side wall of the spacer oxide, the third oxide layer O_3 serving as a blocking oxide.
20. A memory structure according to claim 19, wherein the ($O_1-N_1-O_2$) stack comprises substantially thin layers of oxide and nitride.
21. A memory structure according to claim 20, wherein each of the first oxide layer O_1 , the first nitride layer N_1 , and the second oxide layer O_2 is less than 3 nm.
22. A method for operating an AND memory array having columns of bandgap engineered SONOS (BE-SONOS) devices, each column of BE-SONOS devices corresponding to a sub-gate line and a wordline, comprising:
- resetting a plurality of BE-SONOS devices in the AND memory array, the plurality of BE-SONOS devices producing a self-converging reset to reset a threshold voltage V_t value;
 - programming electrically a selected BE-SONOS device in a first column of the BE-SONOS devices; the first column of the BE-SONOS devices corresponding to a first wordline which is applied a high voltage, the first column of the BE-SONOS devices corresponding to a first sub-gate line which is applied a low voltage to perform source side injection; and
 - erasing electrically the plurality of BE-SONOS devices in the AND memory array, the plurality of BE-SONOS devices producing a self-converging erase to the reset voltage threshold V_t value.
23. A method according to claim 22, further comprising reading the selected BE-SONOS device, wherein the voltage of the first wordline is raised to a voltage that is between an erases state level and a program state level.
24. A method according to claim 22, wherein the first sub-gate line serving as source side injection during

programming when the first sub-gate line is a slightly turn-on state.

- 25.** A method according to claim 22, wherein the first sub-gate line serving as an inversion layer for connection to a first metal bitline when the first sub-gate line is an ON state. 5
- 26.** A method according to claim 22, wherein the low voltage applied to the first sub-gate line is 1-2 volt. 10

15

20

25

30

35

40

45

50

55

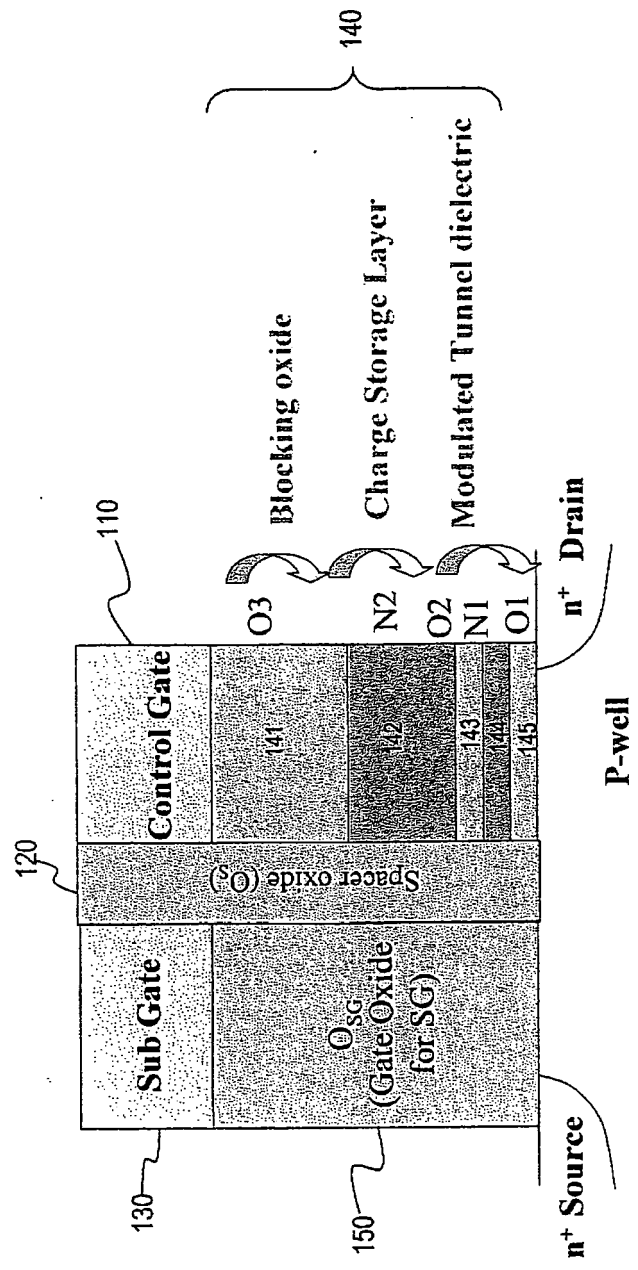


Fig. 1

100

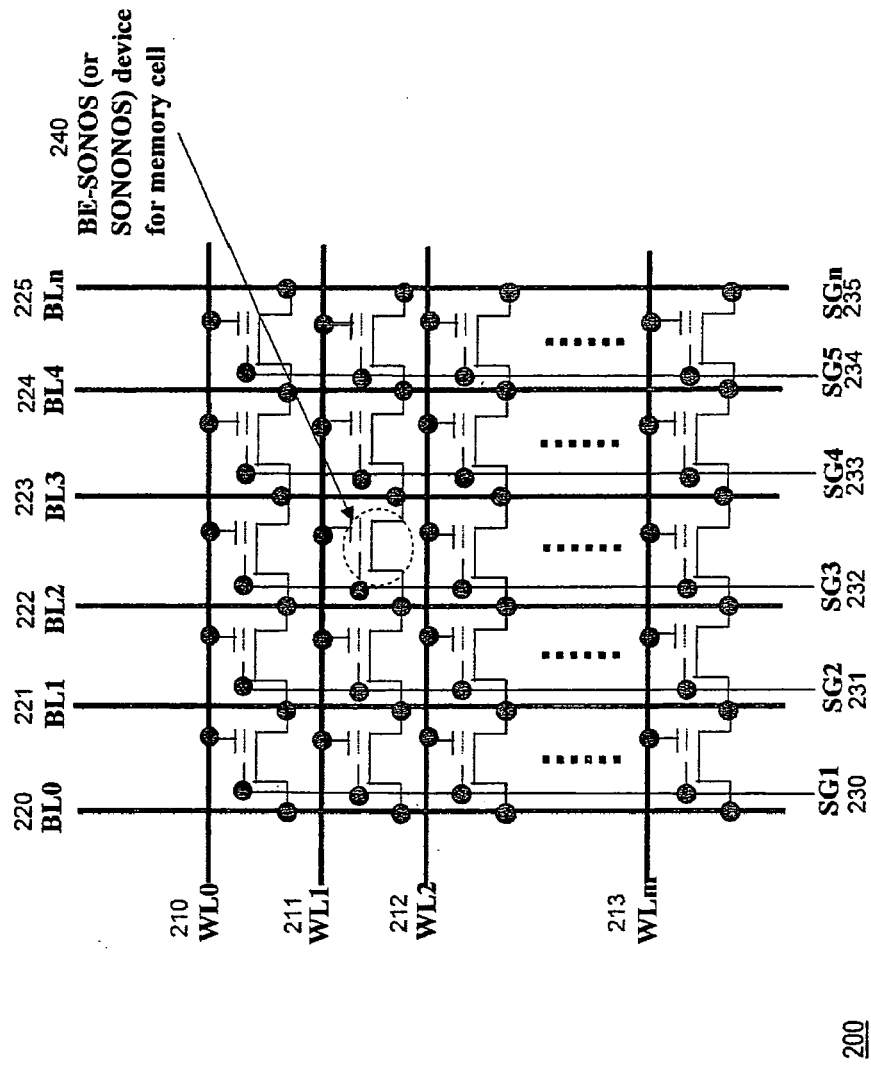


Fig. 2A

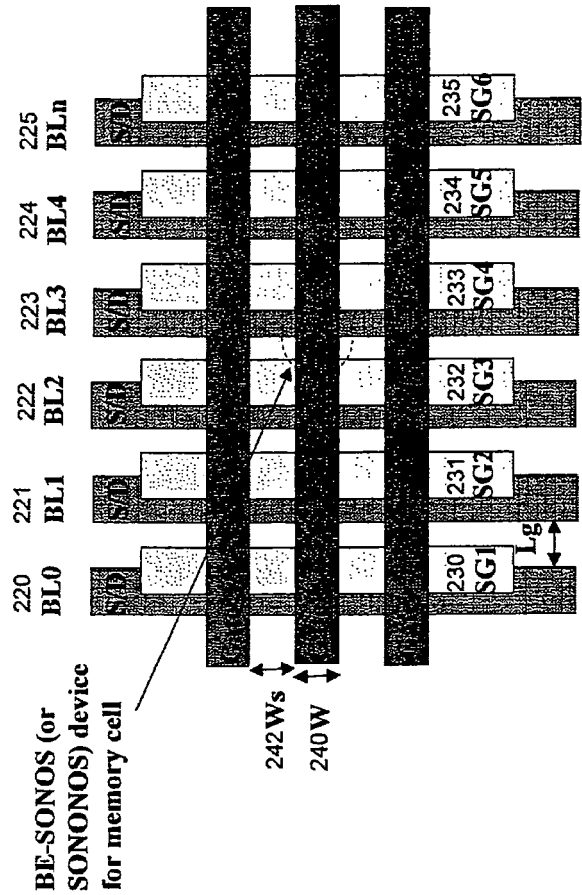


Fig. 2B

250

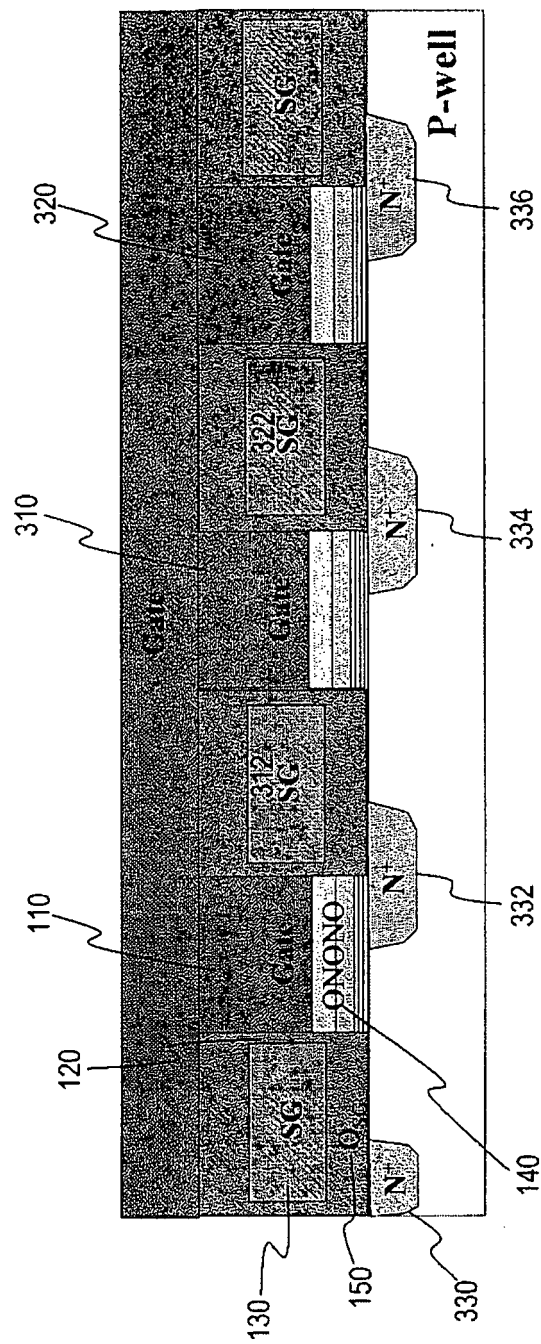
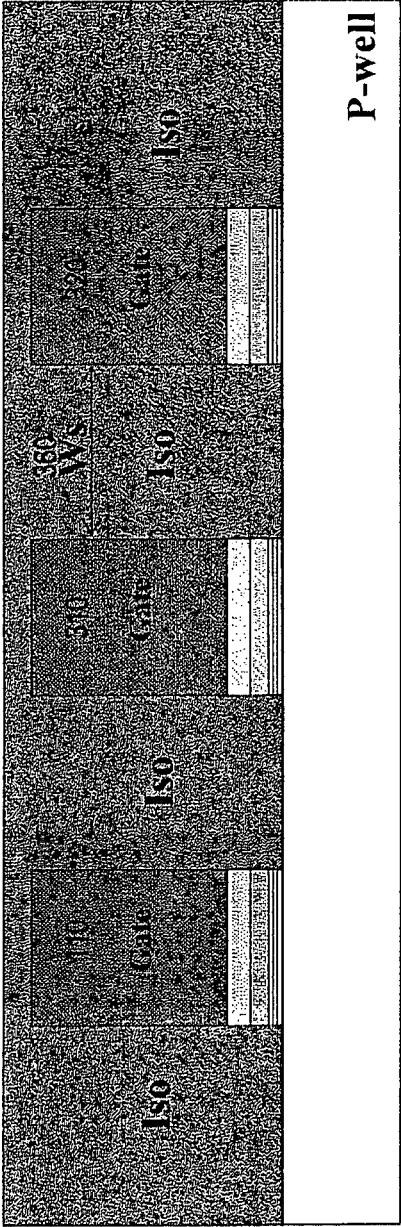


Fig. 3A

300



350

Fig. 3B

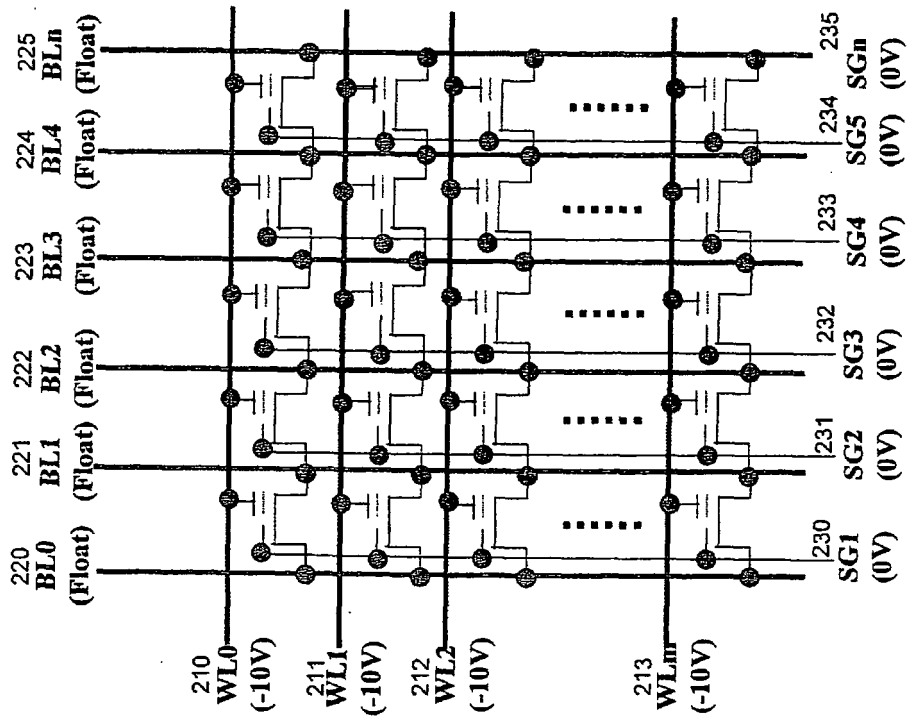


Fig. 4A

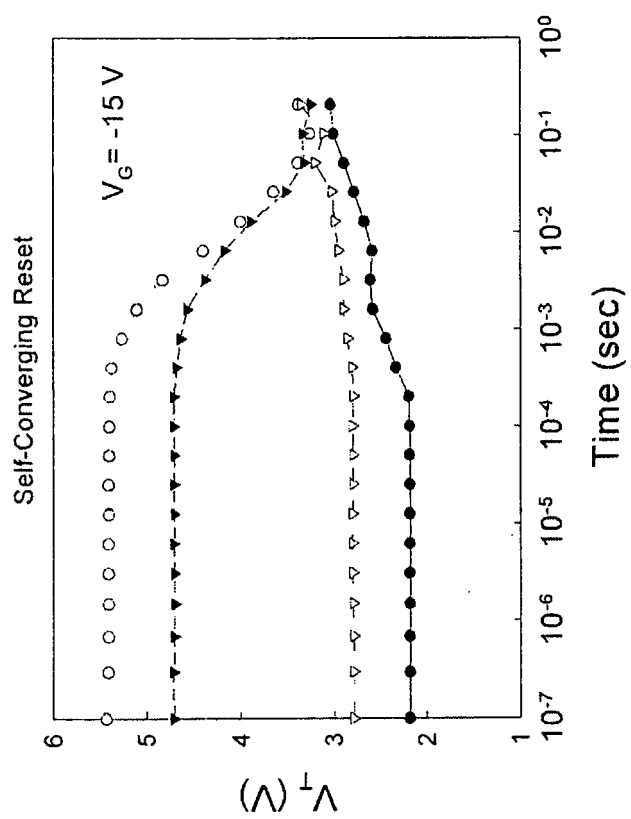


Fig. 4B

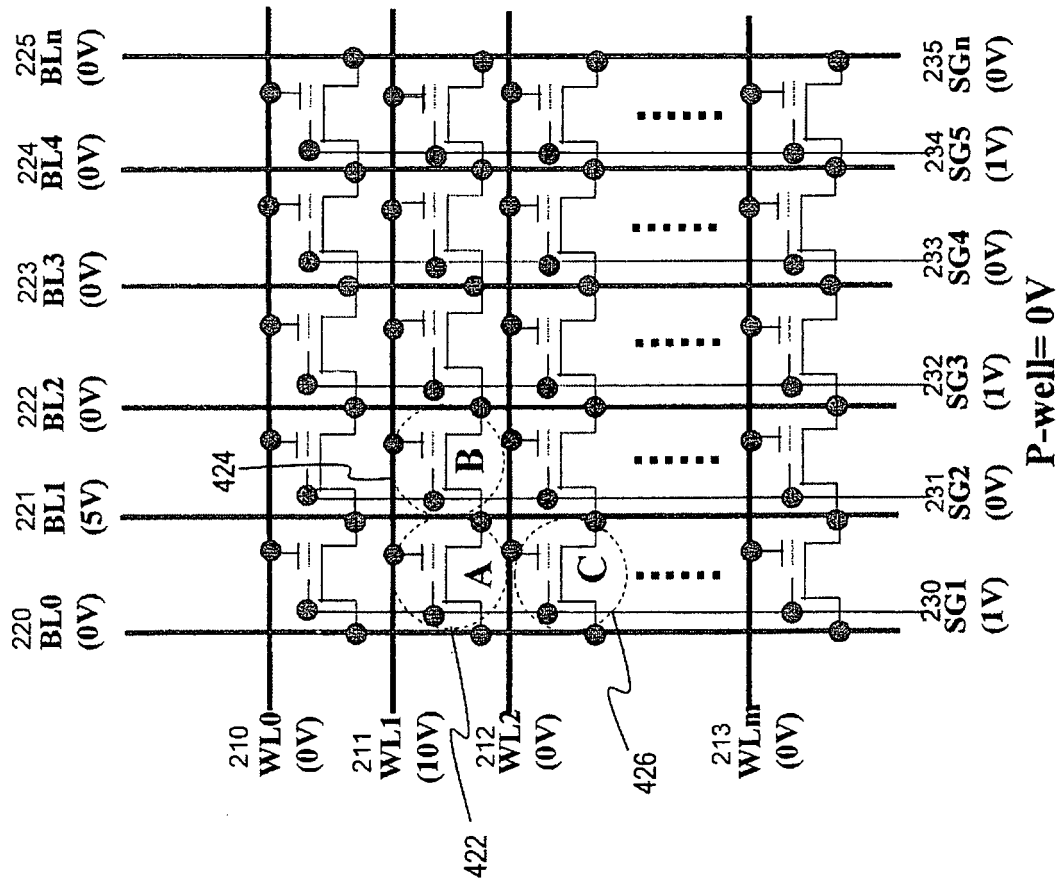


Fig. 5A

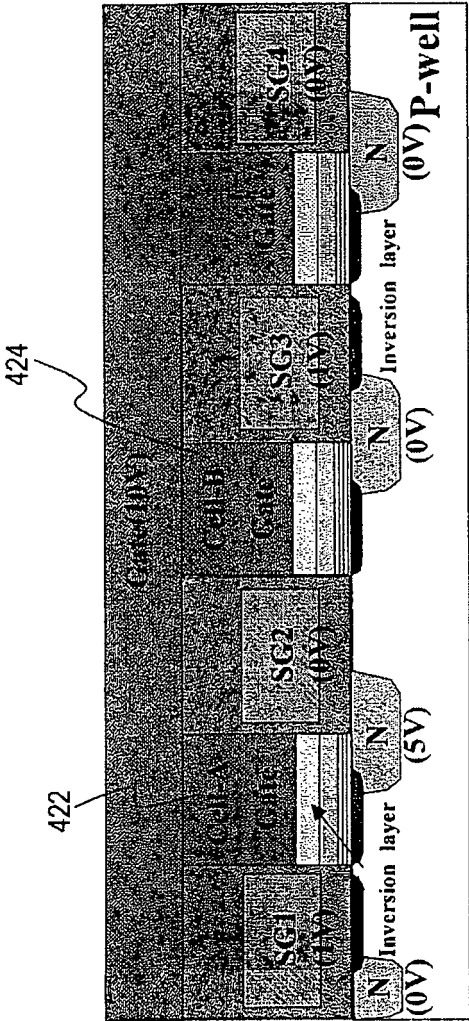


Fig. 5B

550

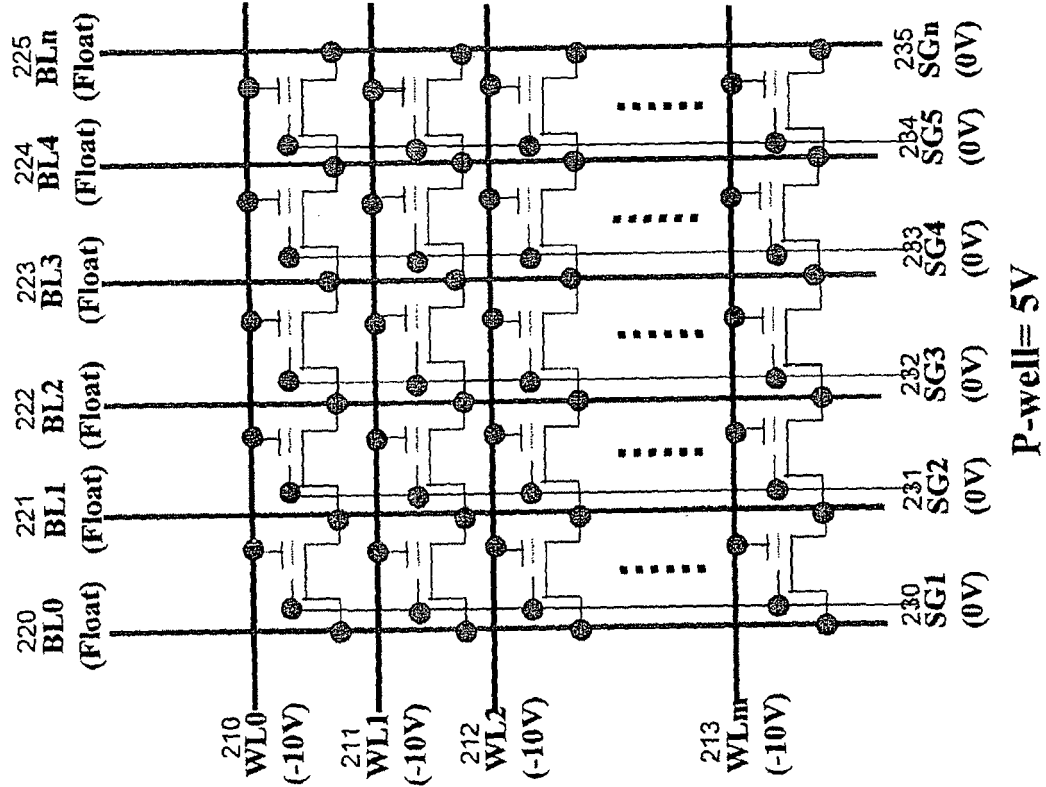


Fig. 6A

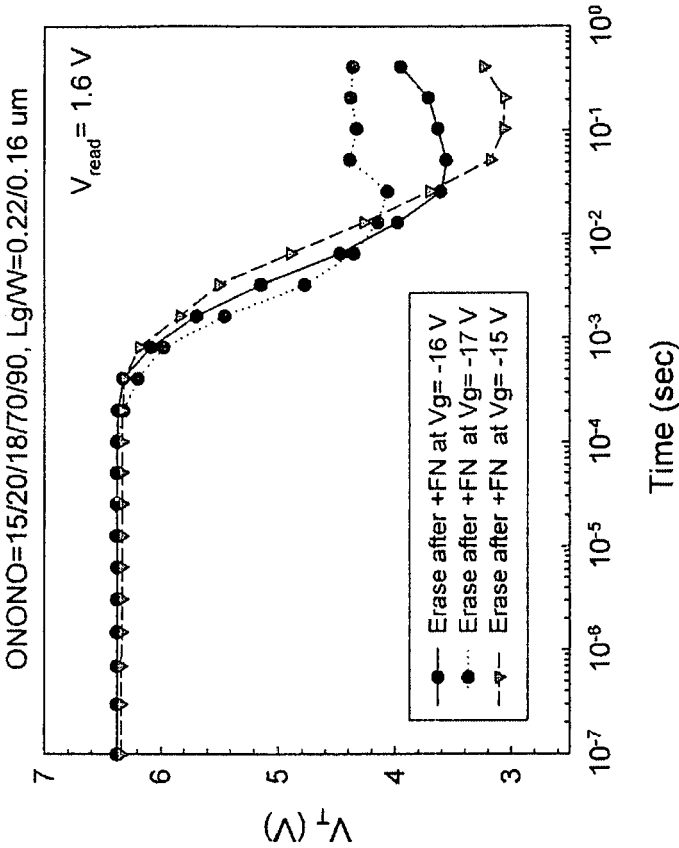


Fig. 6B

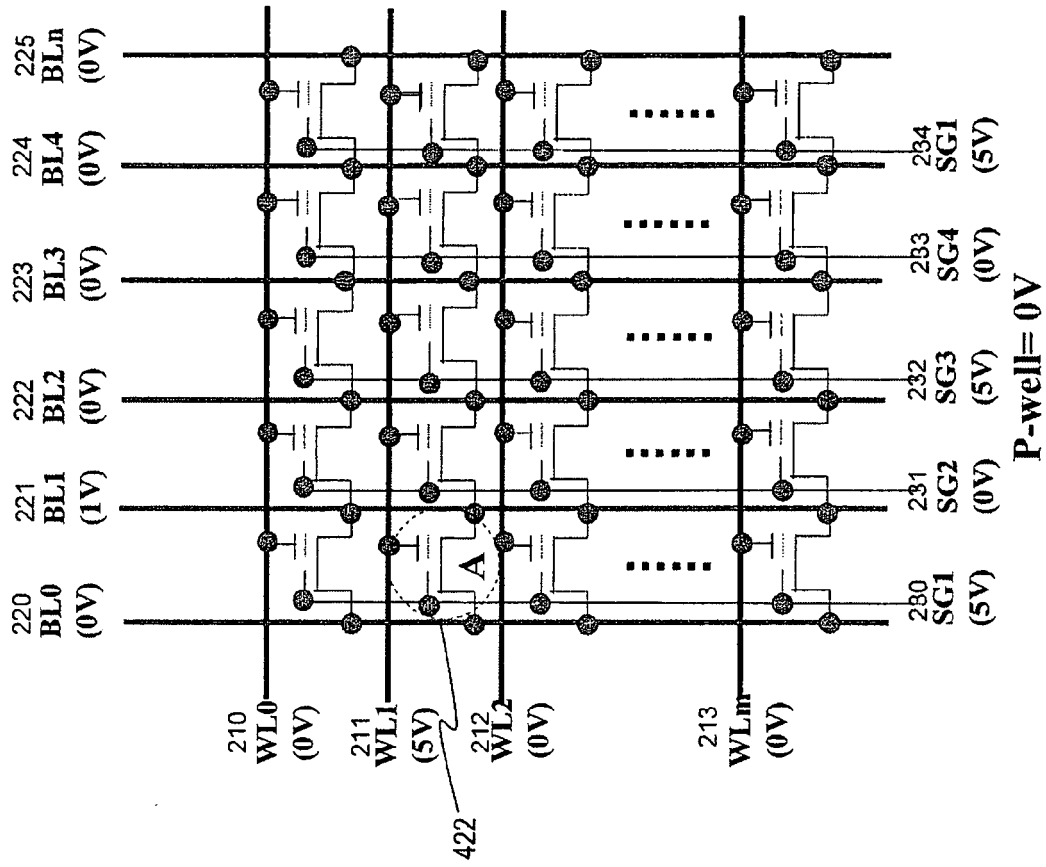


Fig. 7

700

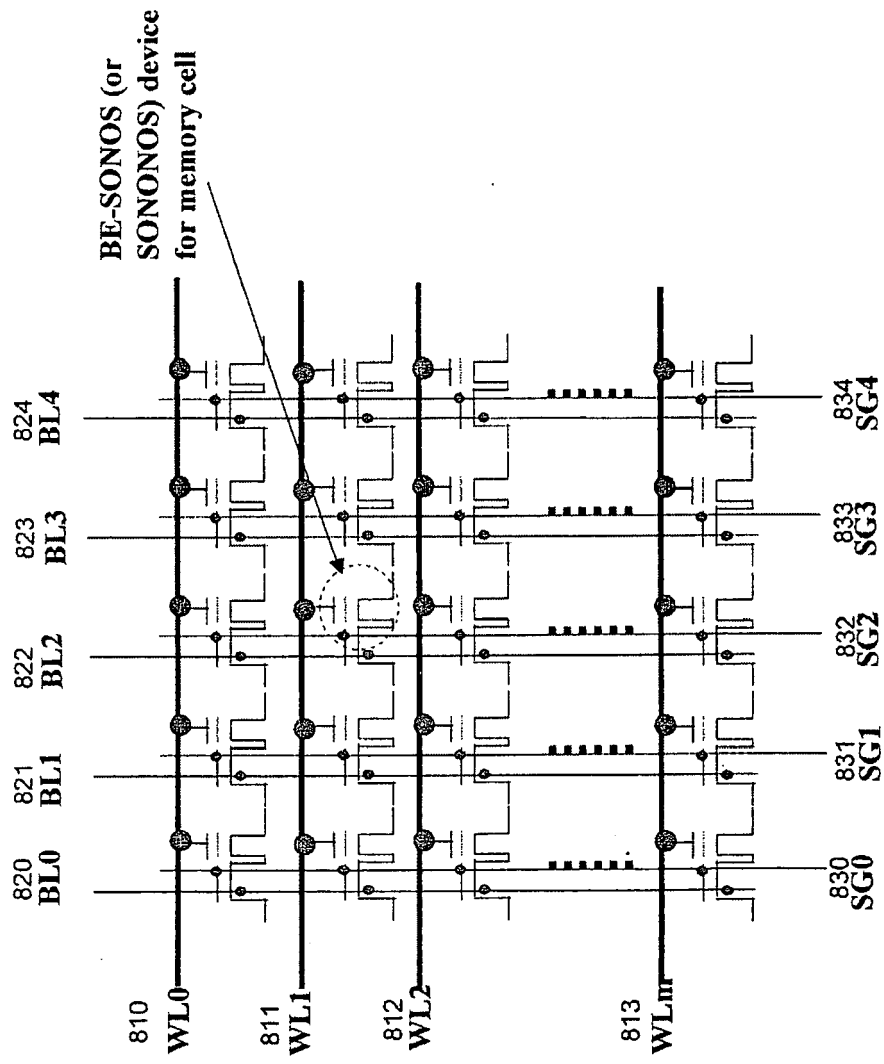
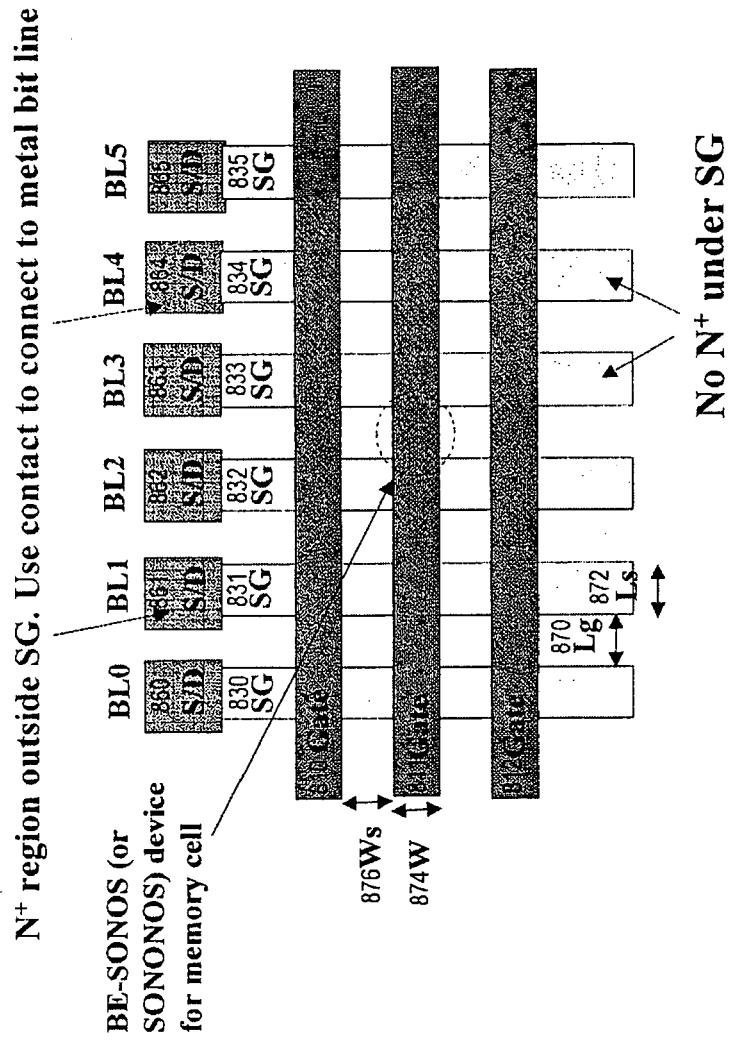


Fig. 8A

800



850

Fig. 8B

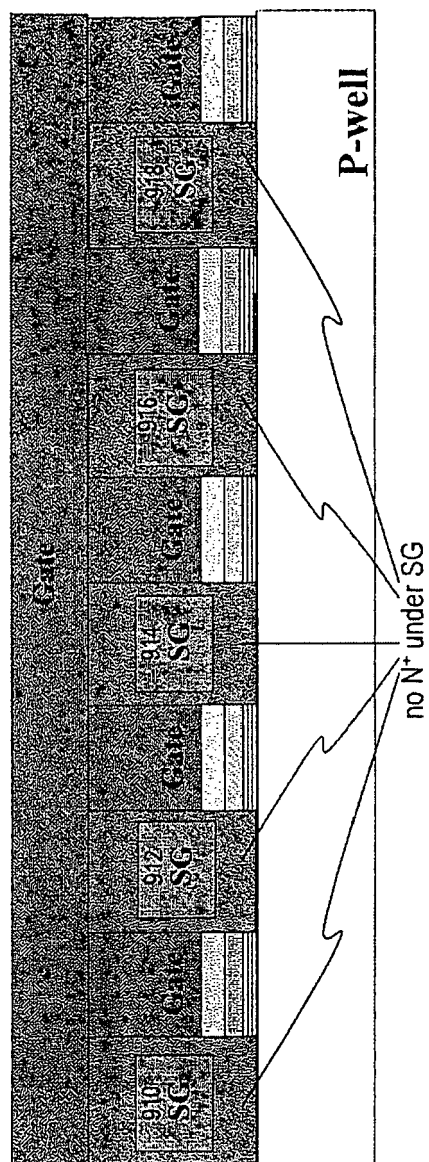


Fig. 9A

900

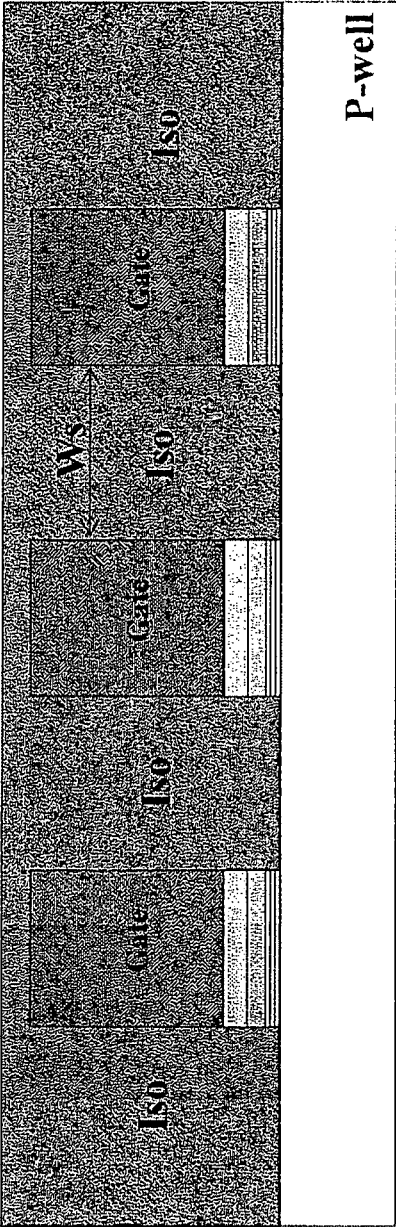
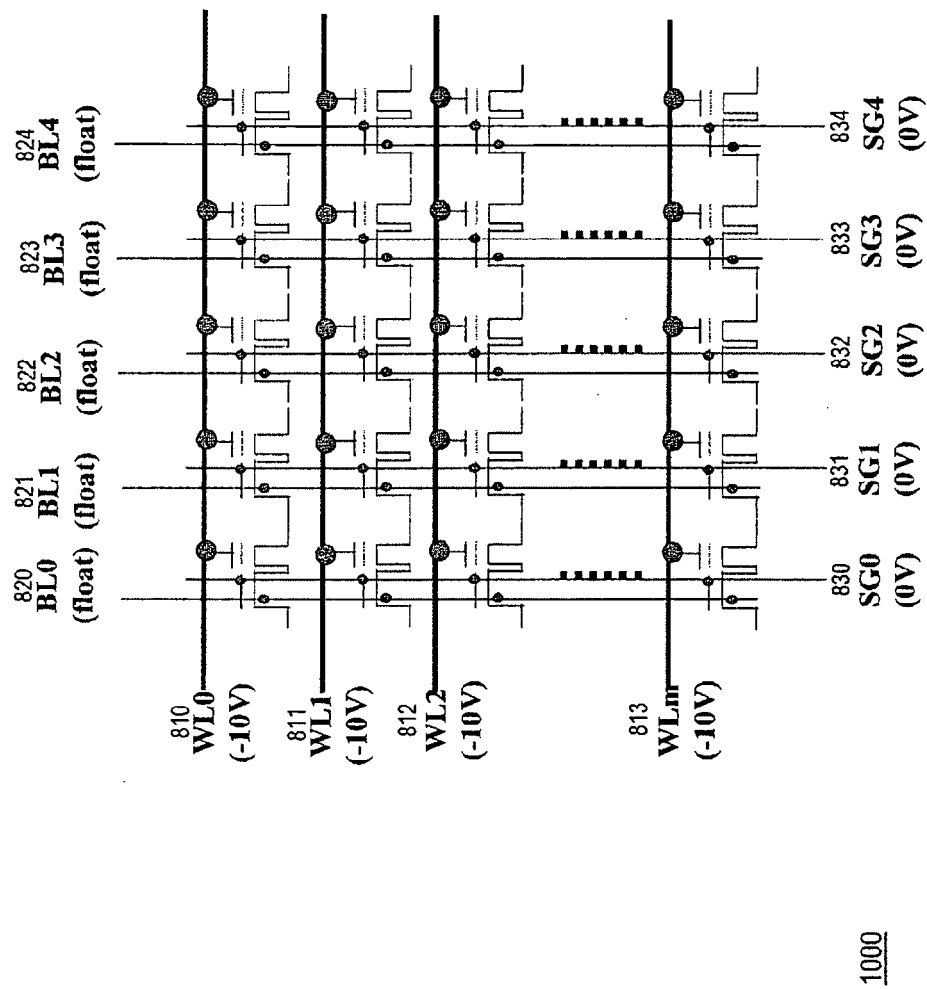


Fig. 9B



¹⁰¹⁰
P-well= 5V

Fig. 10A

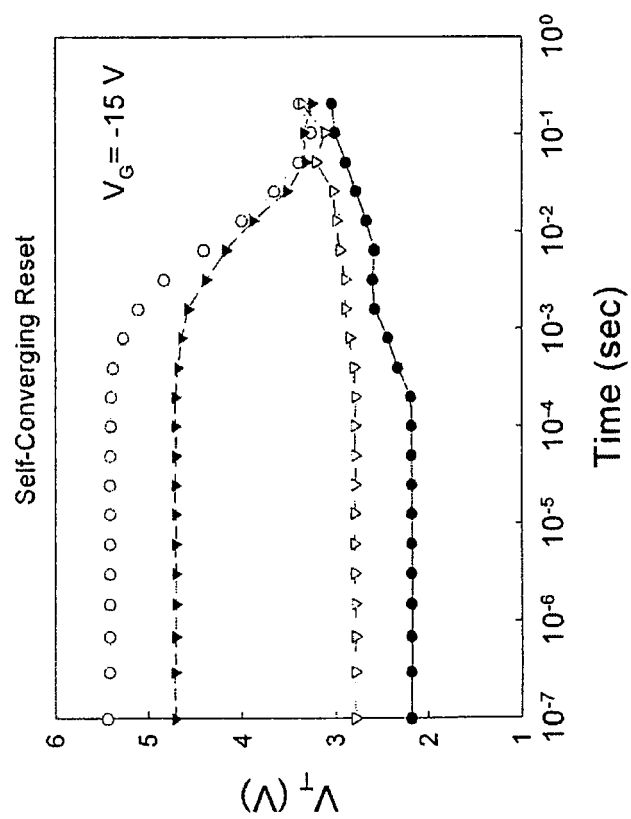


Fig. 10B

1050

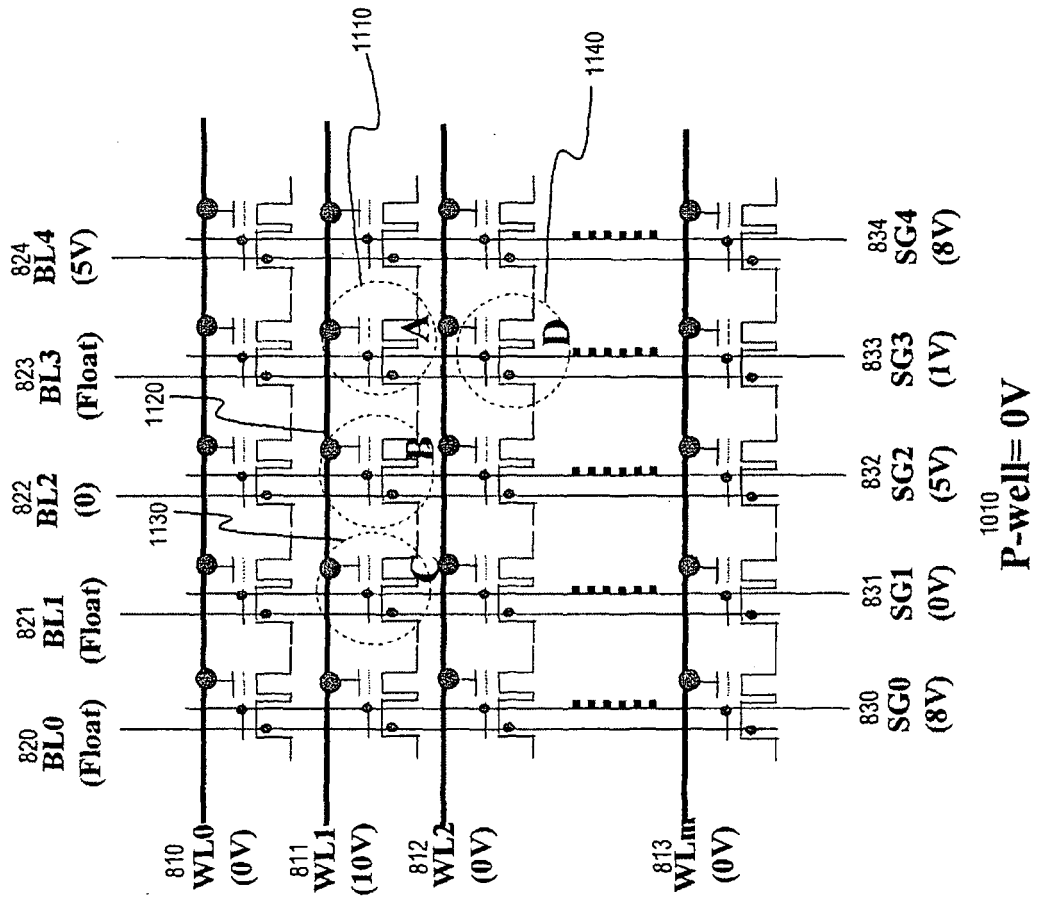


Fig. 11A

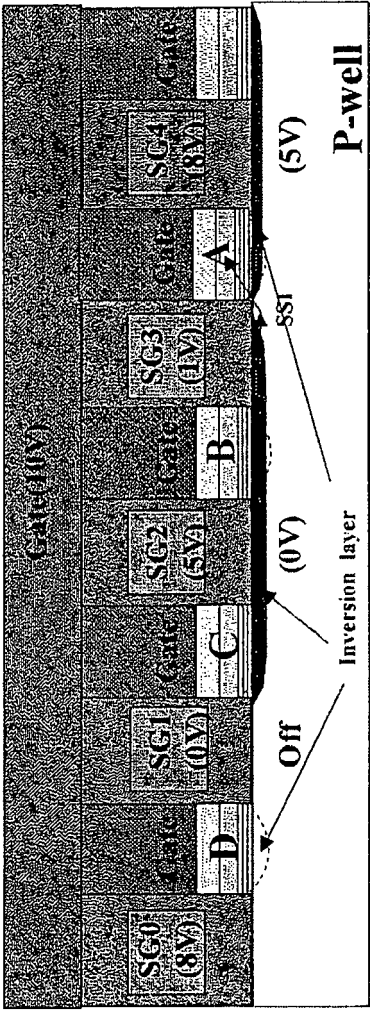


Fig. 11B

1150

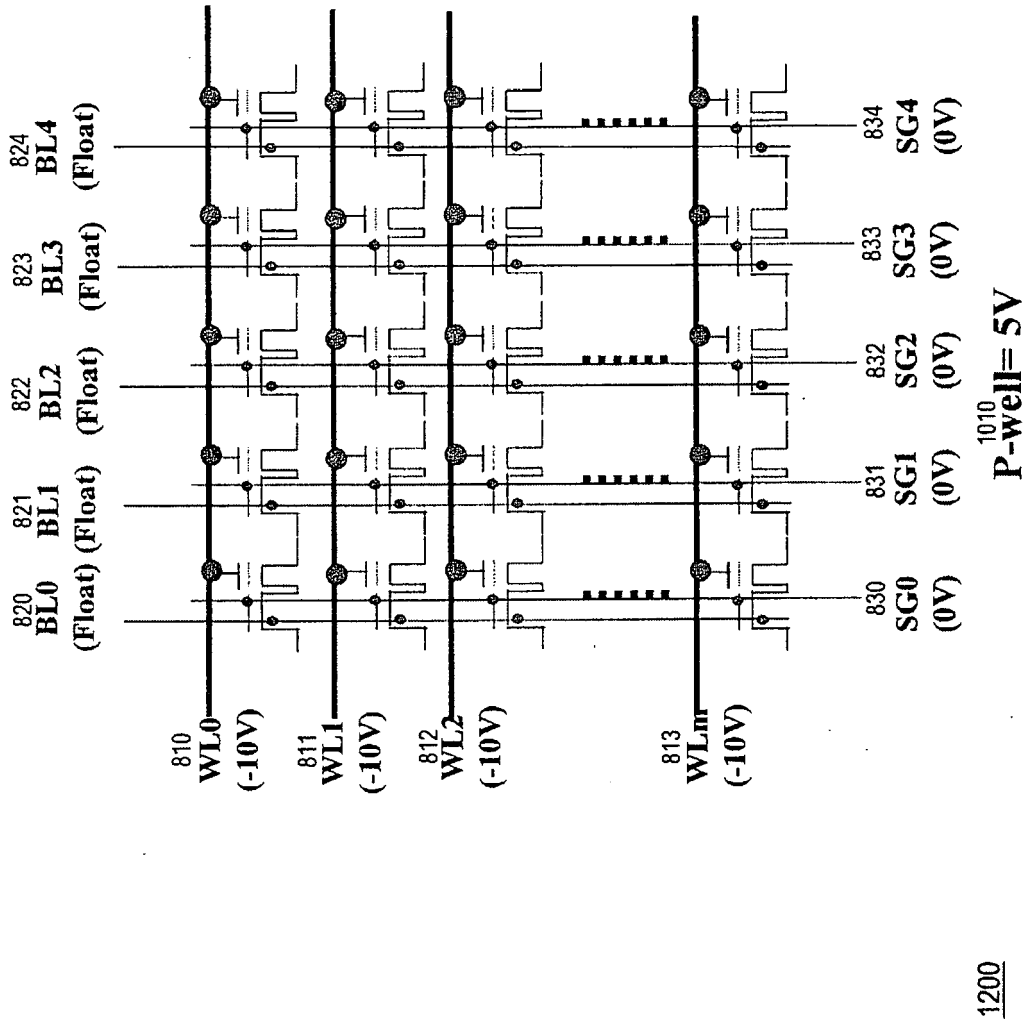
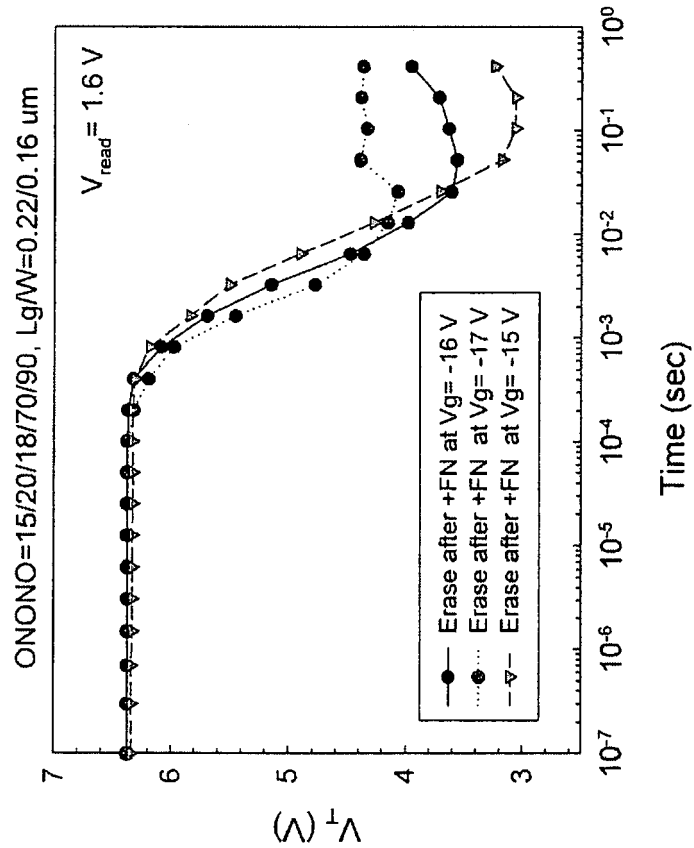


Fig. 12A



1250

Fig. 12B

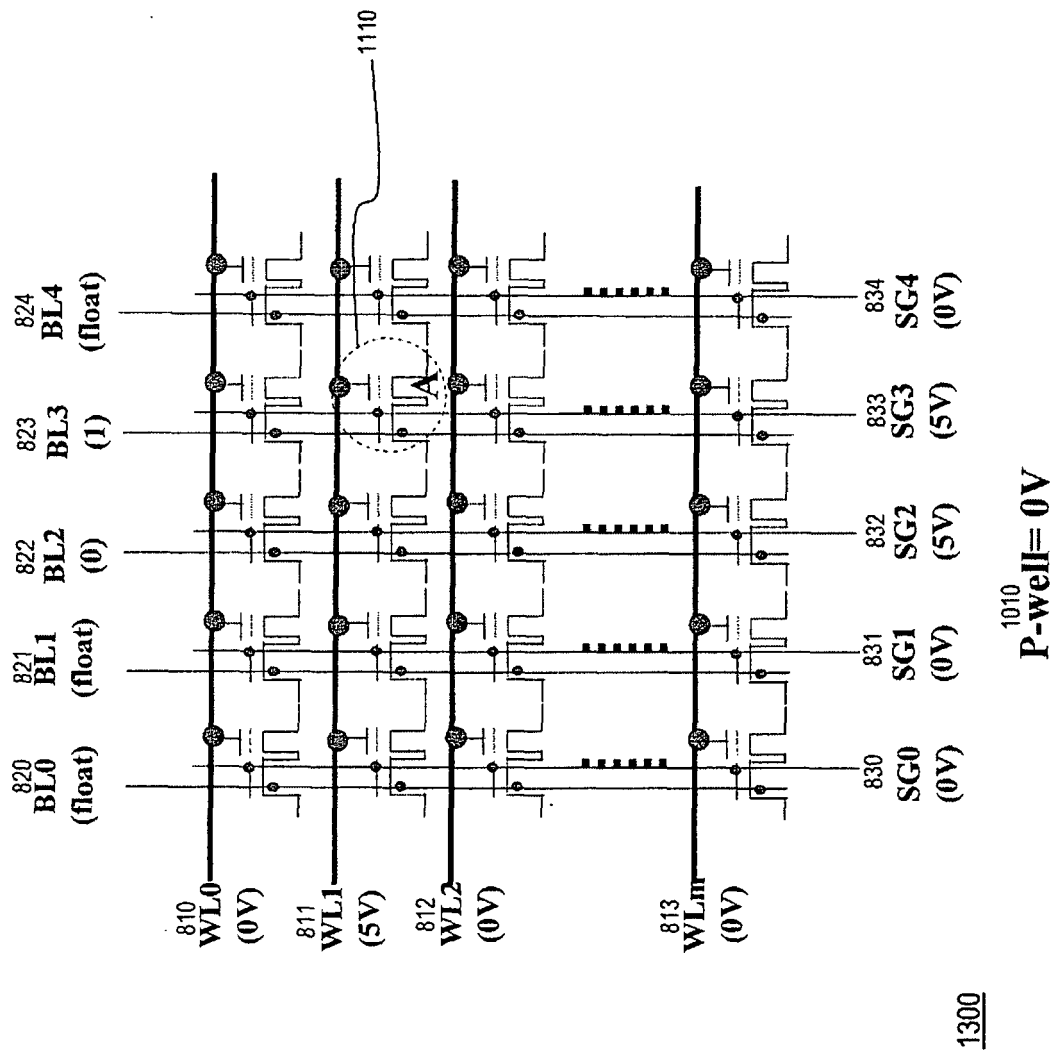


Fig. 13A

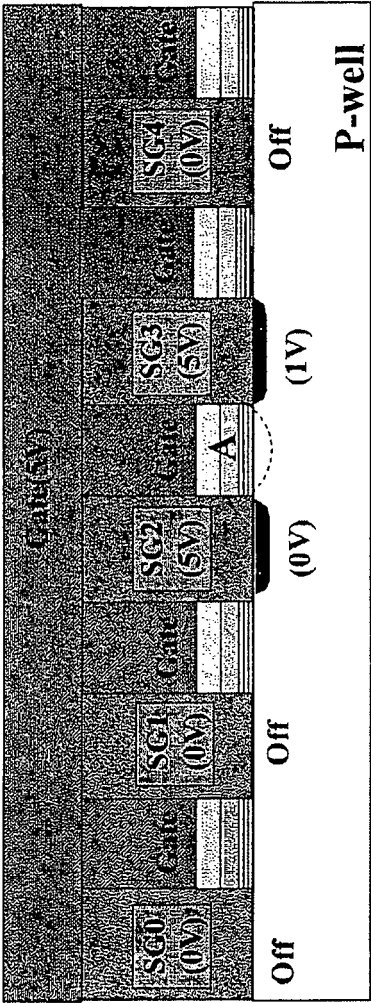


Fig. 13B

1350